CHIWEI: A code of goodness of fit tests for weighted and unweighted histograms

N.D. Gagunashvili[a,*]

[a] *University of Akureyri, Borgir, v/Nordurslód, IS-600 Akureyri, Iceland*

Abstract

A Fortran-77 program for goodness of fit tests for histograms with weighted entries as well as with unweighted entries is presented. The code calculates test statistics for case of histogram with normalized weights of events and in case of unnormalized weights of events.

Keywords: chi-square test generalization, comparison experimental and simulated data, data interpretation, Monte Carlo method

PROGRAM SUMMARY

Program Title: CHIWEI
Journal Reference:
Catalogue identifier:
Licensing provisions: none
Programming language: Fortran-77
Computer: Any Unix/Linux workstation or PC with a Fortran-77 compiler
Classification: 4.13, 11.9, 16.4, 19.4
External routines/libraries used: FPLSOR (M103) from CERN Program Library
Nature of problem: The program calculates goodness of fit test statistics for weighted histograms
Solution method: Calculation of test statistics is done according formulas presented in Ref. [1]

[*] Corresponding author.
E-mail address: nikolai@simnet.is

Preprint submitted to Computer Physics Communications *January 19, 2013*

References

[1] N.G. Gagunashvili, Nucl. Instrum. Meth. A596 (2008) 439.

1. Introduction

A histogram with m bins for a given probability density function $p(x)$ is used to estimate the probabilities

$$p_i = \int_{S_i} p(x)dx, \; i = 1, \ldots, m \tag{1}$$

that a random event belongs to bin i. Integration in (1) is done over the bin S_i.

A histogram can be obtained as a result of a random experiment with probability density function $p(x)$. Let us denote the number of random events belonging to the ith bin of the histogram as n_i. The total number of events in the histogram is equal to $n = \sum_{i=1}^{m} n_i$. The quantity $\hat{p}_i = n_i/n$ is an estimator of p_i with expectation value $\mathrm{E}\,\hat{p}_i = p_i$.

The problem of goodness of fit is to test the hypothesis

$$H_0 : p_1 = p_{10}, \ldots, p_{m-1} = p_{m-1,0} \text{ vs. } H_a : p_i \neq p_{i0} \text{ for some } i, \tag{2}$$

where p_{i0} are specified probabilities, and $\sum_{i=1}^{m} p_{i0} = 1$. The test is used in a data analysis for comparison theoretical frequencies np_{i0} with the observed frequencies n_i. The test statistic

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - np_{i0})^2}{np_{i0}} \tag{3}$$

was suggested by Pearson [2]. Pearson showed that the statistic (3) has approximately a χ^2_{m-1} distribution if the hypothesis H_0 is true.

To define a weighted histogram let us write the probability p_i (1) for a given probability density function $p(x)$ in the form

$$p_i = \int_{S_i} p(x)dx = \int_{S_i} w(x)g(x)dx, \tag{4}$$

where

$$w(x) = p(x)/g(x) \tag{5}$$

is the weight function and $g(x)$ is some other probability density function. The function $g(x)$ must be > 0 for points x, where $p(x) \neq 0$. The weight $w(x) = 0$ if $p(x) = 0$, see Ref. [3]. Because of the condition $\sum_i p_i = 1$ further we will call the above defined weights normalized weights as opposite to the unnormalized weights $\breve{w}(x)$ which are $\breve{w}(x) = const \cdot w(x)$.

The histogram with normalized weights was obtained from a random experiment with a probability density function $g(x)$, and the weights of the events were calculated according to (5). Let us denote the total sum of the weights of the events in the ith bin of the histogram as

$$W_i = \sum_{k=1}^{n_i} w_i(k) \tag{6}$$

and the total sum of squares of weights as

$$W_{2i} = \sum_{k=1}^{n_i} w_i(k)^2, \tag{7}$$

where n_i is the number of events at bin i and $w_i(k)$ is the weight of the kth event in the ith bin. The total number of events in the histogram is equal to $n = \sum_{i=1}^{m} n_i$, where m is the number of bins. The quantity $\hat{p}_i = W_i/n$ for the histogram with normalized weights is the estimator of p_i with the expectation value $\mathrm{E}\,\hat{p}_i = p_i$. Note that in the case where $g(x) = p(x)$, the weights of the events are equal to 1 and the histogram with normalized weights is the usual histogram with unweighted entries.

For weighted histograms again the problem of goodness of fit is to test the hypothesis

$$H_0 : p_1 = p_{10}, \ldots, p_{m-1} = p_{m-1,0} \text{ vs. } H_a : p_i \neq p_{i0} \text{ for some } i, \tag{8}$$

where p_{i0} are specified probabilities, and $\sum_{i=1}^{m} p_{i0} = 1$.

The test statistic that is a generalization of Pearson's statistic (3) was proposed in [1] for cases of histograms with normalized weights of entries as well as with unnormalised weights of entries. A code for the calculation of test statistics is presented in this article. As shown in [1] if hypothesis H_0 (8) is true then the statistic for a histogram with normalized weighted entries has approximately the χ^2_{m-1} distribution and for a histogram with unnormalized weighted entries has χ^2_{m-2} distribution.

Use of the proposed test is inappropriate if any expected count in bin of histogram is below 1 or if the expected count is less than 5 in more than 20% of the bins. This empirical restriction known for the usual chi-square test [4] is quite reasonable for weighted histograms.

Information for readers. Recently, another paper dedicated to weighted histograms has been published in ”Computer Physics Communication“, see Ref. [6]. The same author has presented a program for calculating test statistics to compare weighted histogram with unweighted histogram and two histograms with weighted entries. The test can be used for the comparison of experimental data distributions with simulated data distributions as well as for the two simulated data distributions.

2. Computer program

CHIWEI is subroutine which can be called from Fortran program for the calculation of test statistics.

Usage

```
CALL CHIWEI(P,W1,W2,N,NCHA,MODE,STAT,NDF,IFAIL)
```

Input Data

P – one dimensional real array of probabilities p_i

W1 – one dimensional array, sum of weights W_i in each bin

W2 – one dimensional array, sum of squares of weights W_{2i} in each bin

N – number of events n

NCHA – number of bins m

MODE – must be equal to 1 for a histogram with normalized weights, and equal 2 for histogram with unnormalized weights

Output data

STAT – test statistic following a chi-square distribution with NDF degrees of freedom if hypothesis H_0 is true

NDF – number of degree of freedom (will be m-MODE)

IFAIL – will be >0 if calculation is not successful.

3. Test run

We take a distribution

$$p(x) \propto \frac{2}{(x-10)^2+1} + \frac{1}{(x-14)^2+1} \tag{9}$$

defined on the interval $[4, 16]$ and representing two so-called Breit-Wigner peaks. Two cases of the probability density function $g(x)$ are considered

$$g_1(x) = p(x) \tag{10}$$

$$g_2(x) \propto \frac{2}{(x-9)^2+1} + \frac{2}{(x-15)^2+1} \tag{11}$$

Distribution (10) gives an unweighted histogram and the method coincides with Pearson's chi square test. Distribution (11) has the same form of parametrization as (9), but with different values of the parameters. Three cases of histograms were considered: unweighted histogram, histogram with weights $p(x)/g_2(x)$ and histogram with unnormalized weights $2p(x)/g_2(x)$. Histograms with 5 bins were created by simulation 1000 entries for each case. The results of the calculations are presented below. Program PROB(G100) [5] has been used for calculating p-values.

Test 1

```
               INPUT

P         0.0296    0.1106    0.4460    0.2067    0.2072
W1       26.0000  115.0000  454.0000  183.0000  222.0000
W2       26.0000  115.0000  454.0000  183.0000  222.0000
```

```
N      1000
NCHA      5
MODE      1

               OUTPUT

STAT      4.5291               (p-value=0.3391)
NDF       4
IFAIL     0
```

Test 2

```
               INPUT

P         0.0296    0.1106    0.4460    0.2067    0.2072
W1       36.0112  106.1355  458.3037  197.8123  205.7211
W2       28.2698   56.9601  938.7897  363.4649  172.2003
N      1000
NCHA      5
MODE      1

                OUTPUT

STAT      2.3380               (p-value=0.6738)
NDF       4
IFAIL     0
```

Test 3

```
                INPUT

P         0.0296    0.1106    0.4460    0.2067    0.2072
W1       72.0225  212.2710  916.6075  395.6246  411.4423
W2      113.0790  227.8403 3755.1587 1453.8595  688.8014
N      1000
NCHA      5
MODE      2

               OUTPUT
```

```
STAT     2.2398                    (p-value=0.5241)
NDF      3
IFAIL    0
```

References

[1] N.G. Gagunashvili, Nucl. Instrum. Meth. A596 (2008) 439.

[2] K. Pearson, Phil. Mag. 5th Ser. 50 (1900) 157.

[3] I. Sobol, A Primer For The Monte Carlo Method, CRC Press, Boca Raton, Florida, 1994.

[4] D.S. Moore, G.P. McCabe, Introduction to the Practice of Statistics, W.H. Freeman Publishing Company, New York, 2005.

[5] CERN Program Library, http://cernlib.web.cern.ch/cernlib/.

[6] N.D. Gagunashvili, Comp. Phys. Comm. CPC-D-11-00196R1